Red Bayesiana a partir de factores de riesgo de la Hipertensión Arterial

Autores: MSc. María del Carmen Chávez Cárdenas (mchavez@uclv.edu.cu)

Lic. Santiago Cuadrado Rodríguez (santiagocr@uclv.edu.cu)

Dra. Gladys Casas Cardoso (gladita@uclv.edu.cu)

MSc. Natalia Martinez Sánchez (natalia@ucv.edu.cu)

Universidad Central de Las Villas (UCLV)

Santa Clara. Villa Clara. CUBA. C.P. 54830.

Tel: (53)-(422)-81515. Fax: (53)-(422)-81608

Resumen

La hipertensión arterial (HTA) es un factor de riesgo para numerosas enfermedades, sin embargo ella por sÍ misma representa una enfermedad. En esta investigación se realizan análisis a partir de distintos factores de riesgos y atributos que permiten caracterizar los tipos de HTA.

A partir de casos con su diagnóstico se propone un método para obtener la topología de una red bayesiana, la cual utiliza la técnica del CHAID (*Chi-squared Automatic Interaction Detector*) como herramienta intermedia y va construyendo *árboles de decisión* que muestran las dependencias entre dichas variables. Para completar el modelo probabilístico que conforma la red, se calculan las probabilidades condicionales con ayuda de los paquetes SPSS (*Statistical Package for Social Science*) y *Mathematica*. La red bayesiana obtenida es útil para realizar tanto el diagnóstico de la Hipertensión arterial, como de analizar cual sería el comportamiento de los atributos predictores más significativos.

Introducción

La hipertensión arterial es un factor de riesgo para las enfermedades del corazón, pero ella por si misma representa una enfermedad. La Organización Mundial de Salud la ha denominado epidemia silenciosa pues por lo regular se presenta de forma asintomática, ocasionando daños como: trombosis, hemorragias cerebrales, infarto del miocardio, muerte súbita, insuficiencia renal, entre otras. En la investigación se realizan análisis a partir de distintos factores de riesgos y atributos que permiten caracterizar los tipos de HTA.

Como parte de un proyecto de investigación conjunta entre la Universidad Central de Las Villas y la Universidad de Oviedo, hace algunos años se creó la “Proyección del Centro de Desarrollo Electrónico hacia la Comunidad” (PROCDEC) cuyo objetivo principal es desarrollar un estudio de personas supuestamente normotensas primero en la ciudad de Santa Clara y luego en toda la nación.

En el desarrollo de este proyecto participa un grupo multidisciplinario formado por un psicólogo, un cardiólogo, un nefrólogo, un genetista, tres fisiólogos, dos clínicos, un médico de laboratorio, dos ingenieros y dos cibernéticos. Participan además especialistas en Medicina Integral General de los centros hospitalarios José Ramón León, Chiqui Gómez y Ramón Pando Ferrer. Estos especialistas realizan la captura de los datos, mientras el grupo multidisciplinario es quien realiza el diagnóstico.

A partir de casos con su diagnóstico se propone un método para obtener la topología de una red bayesiana la cual utiliza la técnica del CHAID (*Chi-squared Automatic Interaction Detector*) como herramienta intermedia y va construyendo *árboles de decisión* que muestran las dependencias entre dichas variables. Se demuestra como se integran estos árboles para formar la topología de la red bayesiana. Para completar el modelo probabilístico que conforma la red, se calculan las probabilidades condicionales con ayuda de los paquetes SPSS y *Mathematica*. [5],[6]

Desarrollo

Redes de Creencia o Bayesianas

Este tipo de redes se define como un grafo que cumple lo siguiente:

- Los nodos de la red están formados por un conjunto de variables aleatorias que denotamos con la letra X o con subíndices X_1, X_2, … ,X_n. En principio estas variables pueden ser rasgos, pero puede ocurrir también que un mismo rasgo tenga que ser descompuesto en varias variables aleatorias. Por ejemplo, si el rasgo es multivaluado y se desean las variables aleatorias dicotómicas.
- Cada par de nodos se conecta entre sí mediante un conjunto de enlaces o flechas. El significado de una flecha que vaya del nodo X al nodo Y es el de que X ejerce una influencia directa sobre Y, en términos de probabilidades esto significa que hay una dependencia condicional de Y respecto a X, esto es que la probabilidad de Y es diferente de la probabilidad de Y dado X .
- Por cada nodo hay una tabla de probabilidad condicional que sirve para cuantificar los efectos de los padres sobre el nodo. Los padres de un nodo son aquellos nodos cuyas flechas apuntan hacia éste.
- El grafo no tiene ciclos dirigidos (por lo tanto es un grafo acíclico dirigido (GAD)). Esto significa que no se presentan ambigüedades en el encadenamiento de probabilidades condicionales por el hecho de influencias directas cíclicas.

Vista la red de creencias como el grafo junto con las tablas de probabilidad condicional, ella puede ser interpretada como una representación de la función de distribución de probabilidad conjunta (DPC) de la clase y de todos los rasgos (variables). El árbol en sí codifica un conjunto de aseveraciones de independencia condicional. Las tablas de probabilidades condicionales completan la caracterización de la distribución conjunta.

El árbol es importante para construir la red en sí. Los valores que aparecen en las tablas de probabilidades condicionales son imprescindibles en el procedimiento de inferencia. Esta representación es a lo que algunos autores llaman I–mapa minimal de la distribución conjunta [4][11]. Una definición formal se puede ver en [4]

En [11] se formula un algoritmo general para la construcción de las redes de creencia.

Para los distintos sistemas de inferencia probabilísta el objetivo principal es el cálculo de la distribución de probabilidad posterior de un conjunto de variables de consulta, en base a determinadas variables de evidencia. En [4] se hace referencia a distintos algoritmos para la propagación de evidencias (actualización de la distribución de probabilidad de las variables de acuerdo a las nuevas evidencias, por ejemplo, en un problema de diagnóstico medico, se necesita calcular la distribución de probabilidad de enfermedades de interés dados los síntomas), dentro de estos algoritmos tenemos: algoritmos de propagación exacta (propagación de evidencia en poliárboles), propagación de evidencias en redes múltiplemente conectadas, propagación usando métodos para unir árboles), algoritmos de propagación de evidencias aproximados basados en métodos de simulación y algoritmos de propagación simbólica.

Se cuenta con la implementación del algoritmo de propagación Join Tree (Unión de árboles) que permite realizar la inferencia en redes múltiplemente conectadas, a este software se le ha denominado BySHell, este método de propagación es el que más se ha utilizado en los distintos softwares publicos en Internet en la temática, La propagación de evidencia se hace en forma eficiente combinando la información proveniente de los distintos subgrafos, a través del paso de mensajes.

Analizando los pasos de la ingeniería del conocimiento y el algoritmo para construir redes de creencia que se dan en [11], una vez definidas las variables que intervienen en el modelo, se definen las relaciones de independencia condicional de estas variables, sin embargo cuando se trabaja con los expertos en ocasiones resulta difícil responder a preguntas como: ¿qué variables están relacionadas?, ¿en qué orden están relacionadas?. Una vez obtenida la topología de la red, pudiera resultar interesante responder: ¿cuáles son las tablas de probabilidades asociadas?. Haciendo un análisis de estos aspectos se utiliza un método para construir la RB a partir de los datos.

Método para construir redes de creencia.

Paso I. Obtención de la topología de la red.

Para obtener la arquitectura de la red aplicamos el paquete de programas CHAID. El es particularmente útil en todos aquellos problemas en que se quiera subdividir una población a partir de una variable dependiente y posibles variables predictoras que cambien esencialmente los valores de la variable dependiente en cada una de las subpoblaciones o segmentos. [5].

Desde esta formulación inicial, se concibió la posibilidad de aplicación a diversos investigaciones de salud. La más típica de ellas, es precisamente en epidemiología, en el estudio de los factores de riesgo asociados a una enfermedad. En tal caso, la variable dependiente puede ser simplemente la

variable que distingue un grupo de enfermos y sanos y las variables predictivas los posibles factores de riesgo.

De las respuestas de este software se obtiene un árbol con las características mencionadas, el brinda la estructura de la red bayesiana con sólo invertir el orden del árbol.

Paso II. Obtención de las tablas de probabilidades.

Para la obtención de las tablas de probabilidad se propone seguir estos pasos:

- Calcular las probabilidades de todas las variables que son nodos en el grafo utilizando el comando FREQUENCIES del paquete estadístico SPSS, previamente deben ser ponderadas las variables utilizando como pesos los valores de las tasas de la variable dependiente.
- Para cada nodo del grafo que tenga mas de un padre unir estas variables padres y volver al paso 1
- Una vez calculadas las distribuciones de probabilidad para las variables individuales, y para las variables unidas (unión de las variables que son padres de un mismo nodo del grafo), utilizar tablas de contingencia simples (tabla que muestra el número de casos que hay en cada combinación de las categorías de dos variables categóricas [7]) para calcular las probabilidades condicionadas.

Cuando se tiene elaborado el modelo de red Bayesiana, para realizar la inferencia se utilizan algoritmos para la propagación de evidencias, en este trabajo se ha utilizado el software BySHell.

La variable dependiente es el diagnostico (Hipertenso, Hiperreactivo vascular y Normotenso). La base de datos cuenta con 36 rasgos predictores (divididos en factores de riesgo y atributos que representan presión). A partir de la unión de árboles de decisión se obtiene el modelo de RB que permite analizar en cada caso. En el presente estudio se hicieron varios análisis para estudiar las influencias de estoa atributos y en que medida permiten caracterizar la HTA, después de hacer varias corridas con los algoritmos del CHAID se decidió trabajar solamente con la presión arterial media, pues esta está altamente correlacionada con las demás mediciones de las presiones y de ahí que siempre queden en la estructura de la red. En el caso de los factores de riesgo el atributo más discriminante es el índice de masa corporal que tanto considerando los diagnósticos hiperreactivos e hipertensos o solamente hipertensos sin hiperreactivos permanece con una influencia directa sobre el diagnóstico HTA, en la estructura de la RB se pueden ver el resto de las variables y la forma en que permanecen en el estudio.

Construcción de los árboles de decisión

Se utiliza la técnica del CHAID como herramienta intermedia para obtener los *árboles de decisión*. Este método permite reducir grandemente el modelo probabilístico, pues tiene en cuenta solamente las interacciones fundamentales de las variables predictivas con las familias o clases conformadas, ejemplo de árbol se aprecian en las Fig.1, en este caso se tiene el árbol más significativo a los efectos de la HTA si solamente se tienen en cuenta los normotensos e hipertensos.

La Base de Datos (BD) relativa a este estudio está formada por un conjunto de 38 atributos predictivos que se obtienen de estudios y entrevistas realizadas a los pacientes y un atributo objetivo (diagnóstico) cuyo valor (hipertenso, normotenso o hipereactivo) se infiere a partir del valor de los atributos predictivos.

De los 38 atributos se mantienen en la RB solamente 20 variables significativas y sus interacciones para el caso en que se consideran los tres diagnósticos y 18 para el caso que solamente se consideran los normotensos e hipertensos.

El orden de significación de las variables en la técnica del CHAID permite ir construyendo los árboles de decisión. Una vez que la variable pertenece al modelo en cualquiera de los árboles ya construidos no se vuelve a introducir, esto permite reducir la complejidad del modelo y evitar que aparezcan ciclos no compatibles con la definición de red bayesiana.

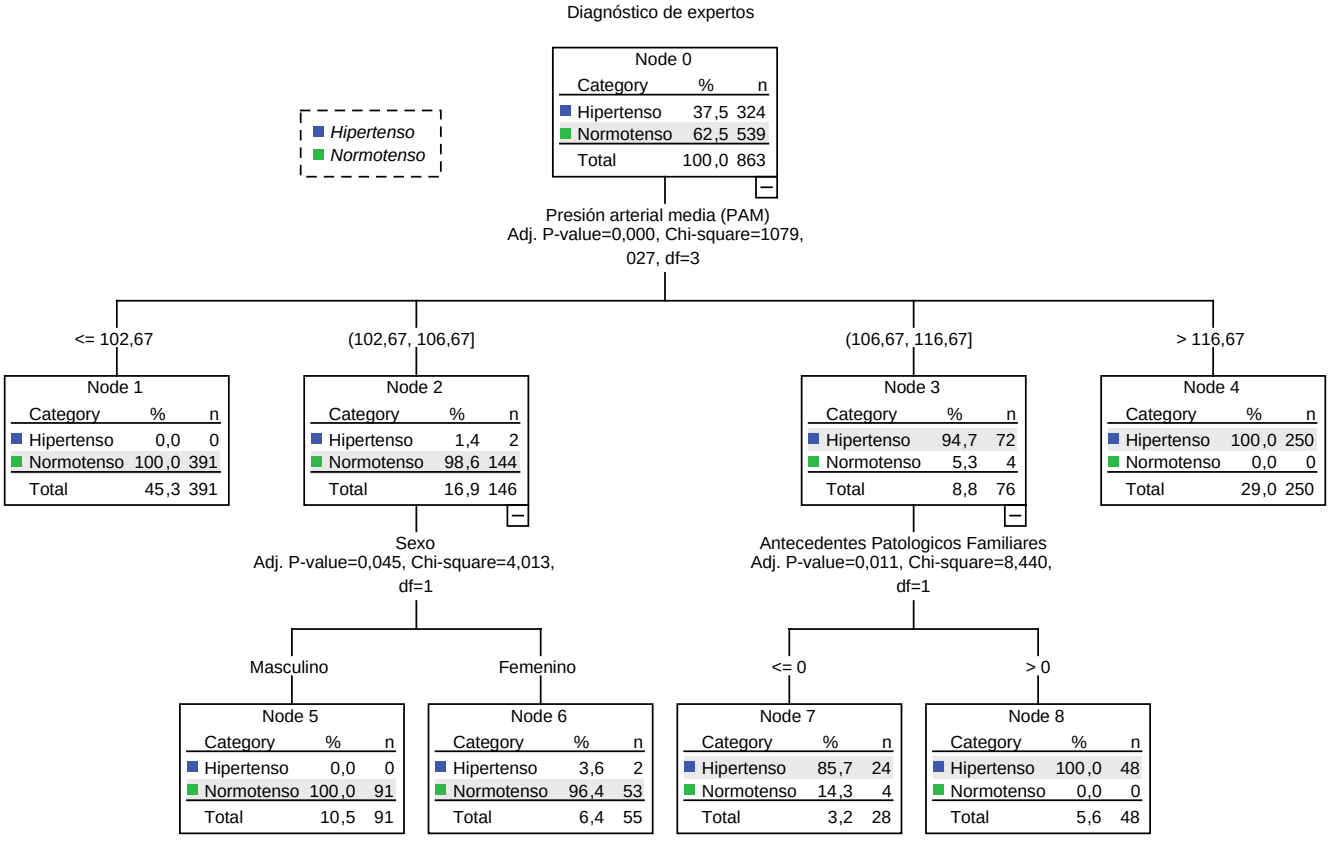

Fig.1 Árbol obtenido para la PAM que es la de mayor interacción con respecto a los tipos de HTA para el caso en que están presentes hipertensos y normotensos.

Construcción de la Red Bayesiana

Los árboles obtenidos establecen una dependencia directa con respecto a la variable dependiente o tipos de HTA. Con la unión de todos los árboles creados se forma el modelo estructural de la red bayesiana, ën este caso se trabaja con un ejemplo en el que aparecen solamente seis atributos significativos, el modelo probabilístico se construye utilizando el método propuesto en [7] en el cual se utiliza el SPSS como herramienta intermedia. [5]

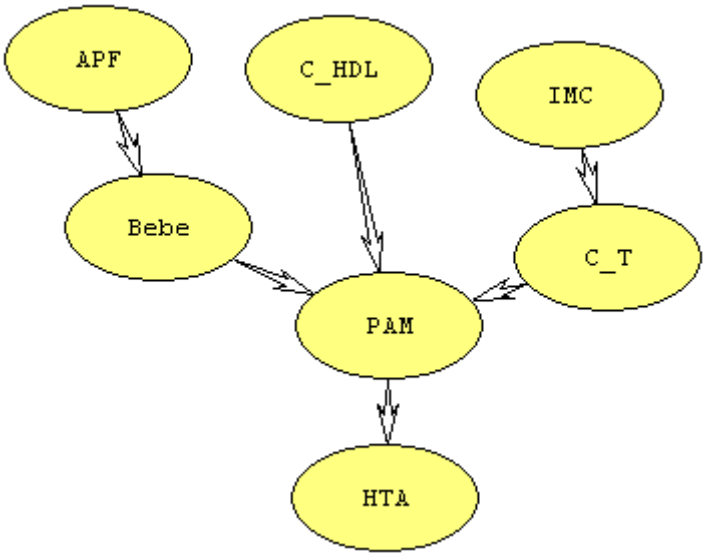

Fig.2 Ejemplo de una *"Red Bayesiana".*

Propagación sobre la red Bayesiana

Se utilizó una primera versión del sofware "ByShell" para propagar evidencias en redes bayesianas. Este sistema tiene implementado el algoritmo de propagación en redes múltiplemente conexas, específicamente el de árboles de conglomerados el cual se reporta como uno de los menos complejos, pero más general para cualquier estructura de RB de los algoritmos de propagación exacta. [4][11]

Con la propagación sobre la red se puede analizar como ante el conocimiento de determinados factores de riesgo cual es el tipo de HTA más probable y que efecto tiene este conocimiento respecto a otros atributos del estudio. En este caso se trabajo con uno de los árboles a modo de RB en la que intervienen los atributos Presión Arterial Media (PAM), Bebe (BB), Colesterol Total (C_T), Colesterol HDL(C_HDL), Antecedentes Patológicos Familiares (APF), Índice de Masa Corporal (IMC).

Por ejemplo si consideramos que un nuevo paciente tiene PAM menor que 93,33 el caso se identifica como normotenso con probabilidad 1(ver fig. 3), si el nuevo paciente tiene PAM mayor a 119 la RB lo identifica como hipertenso con probabilidad 0,97 (ver fig. 4) si el nuevo paciente bebe, y tiene C_T menor igual a 5,4 y C_HDL menor igual a 1,33, la RB lo identifica como normotenso con mayor probabilidad que hipertenso o hiperreactivo, esto se debe a que la PAM es la que tiene una incidencia directa sobre la HTA, según los datos basta con el conocimiento de este atributo para caracterizar la

HTA, en el trabajo se analizaron redes con los atributos considerados factores de riesgo solamente, en estos casos se pueden hacer RB que deben caracterizar a los pacientes acorde a los mismos con mayor credibilidad.

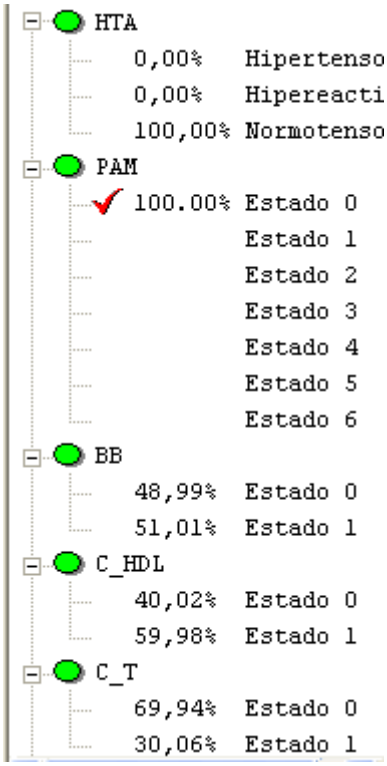

Fig.3 Caso de un paciente normotenso

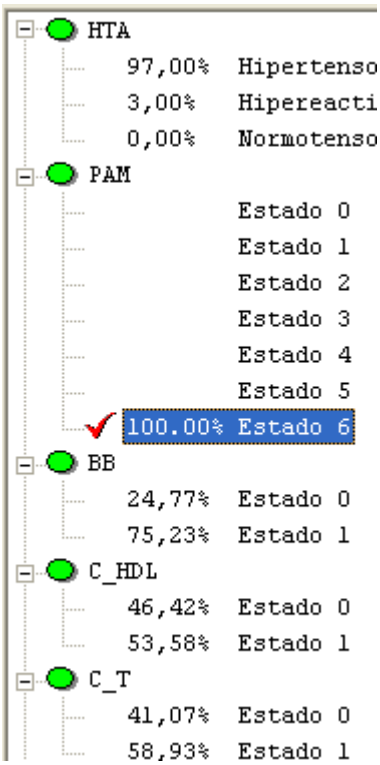

Fig.4 Caso de un paciente Hipertenso

Conclusiones

Se consideran resultados del trabajo la red bayesiana obtenida y los distintos análisis que se pueden hacer con respecto al estudio realizado haciendo uso de la misma.

El ejemplo de la RB garantiza caracterizar pacientes con distintos valores de la PAM, así como ver como se comportan los distintos factores de riesgo que forman parte de la RB.

Bibliografía

[1] Baldi P.,Brunak S., Assessing the accuracy of prediction Algorithms for classification: An Overview. Bioinformatics Vol 16 No. 5 pages 412 – 424, 2000

[2] Barash Y., Friedman N. Context – Specific Bayesian Clsutering for gene expression Data. Annual Conference on Research in Computational Molecular Biology. Pages 12 – 21 ISBN 1 – 58113-353-7. 2001

[3] Brazma A., Jonassen I., Context - especific independence in Bayesian networks, Inc Proc. Twelfth Conference on Uncertainty in Artificial Intelligence, pages 115 – 123, 1996

[4] Castillo Enrique, Gutiérrez J. Manuel, Hadi Ali S. “Expert Systems and Probabilistics Network Models ”, 1996.

[5] CHAID para SPSS sobre Windows. Técnicas de segmentación basadas en razones de verosimilitud Chi-cuadrado “, Manual de usuario, SPSS Soft. Inc. 1994.

[6] Chávez María del C.,”Sistemas de Inferencia probabilística “, Tesis de Maestría 1999.

[7] Jobson, J. D. " Applied Multivariate Data Analysis ". Vol. II: Categorical and Multivariate Methods, Springer, New York, 11-54, 1992

[8] Joseph Bockhorst, Mark Craven, David Page, Jude Shavlik, Jeremy Glasner. "A Bayesian network approach to operon prediction" BIOINFORMATICS Vol. 19 no. 10 2003, pages 1227–1235, 2002

[9] Lam W. " Bayesian network refinement via machine learning approach ". IEEE Trans Patt Anal Mach Int. Vol 20, No.3, 240-251, 1998.

[10] Rodríguez L.O.,Chávez M. C. and. Bayshell, Software para crear redes bayesianas e inferir evidencias en la misma, Registro de Software CENDA, 09358-9358, mayo, 2002

[11] Stuart Rusell, Naving P. "Inteligencia Artificial: Un enfoque Moderno". 1996.

[12] Segal E., Taskar B., Koller D., Probabilistic Classification and Clustering in Relational Data. Inc. Proc. 17th Inter Joint Conference on Artificial Intelligence, 2001

[13] Williams W.L., Wilson,R.C., HancockE.R. " Multiple graph matching with Bayesian inference ". Pattern Recognition Lett. Vol 38, 11-13, 1998